1. ¿Qué ventajas presenta el cálculo de la *Raíz Cuadrada del Error Cuadrático Medio* (RMSE) y su interpretación en relación con su versión ‘inicial’ *Error Cuadrático Medio* (MSE) en problemas de regresión (estimación de variables continuas)?

Rta:

El RMSE y el MSE (son métricas utilizadas para evaluar la precisión de modelos de regresión. Ambas miden el error promedio entre las predicciones del modelo y los valores reales, pero la RMSE tiene ventajas claras sobre el MSE en términos de interpretación y aplicación práctica.

Ventajas principales del RMSE:

* Unidad interpretable: El RMSE está en la misma escala que los valores reales, lo que facilita entender qué tan grande es el error típico de las predicciones.
* Mejor comunicación: Permite expresar el error en términos que tienen sentido para los usuarios.
* Mayor penalización a errores grandes: Al igual que el MSE, el RMSE penaliza más los errores grandes, lo que lo hace útil cuando se desea evitar predicciones con errores extremos.

1. Explique y contraste la diferencia entre procesos de Aprendizaje Supervisado y Aprendizaje No Supervisado. Incluya la explicación de al menos un algoritmo de ejemplo en cada categoría de aprendizaje.

Rta:

* Aprendizaje Supervisad: Es un tipo de aprendizaje donde el modelo se entrena con un conjunto de datos etiquetado, es decir, cada ejemplo de entrenamiento incluye tanto las características (inputs) como la respuesta esperada (output o etiqueta).
  + Ejemplos de algoritmo: Regresion lineal
* Aprendizaje No Supervisado: Es un tipo de aprendizaje donde el modelo se entrena con datos no etiquetados, es decir, no se proporcionan salidas esperadas, y el algoritmo debe descubrir estructuras, patrones o relaciones en los datos por sí solo
  + Ejemplos de algoritmo: K-Means Clustering

1. Asuma un problema de clasificación binaria.
   1. Un algoritmo entrenado con un nivel de prevalencia del 3% que presenta una Precisión del 6% ¿es un mal clasificador? [*Hint: Omita el criterio hipotético con el que se pudo haber seleccionado el punto de corte para calcular la precisión*]

Rta:

Un 6% de precisión implica que, de todas las predicciones positivas que hace el modelo, solo el 6% son realmente verdaderas. Dado que la prevalencia es del 3%, es decir, los casos positivos reales son muy pocos, un clasificador que apenas mejora esa proporción (6%) está cometiendo muchos falsos positivos, lo que indica un alto nivel de ruido en las predicciones positivas. Por lo tanto, este modelo no es confiable en contextos donde identificar correctamente los casos positivos es crítico (por ejemplo, enfermedades raras o fraude)

* 1. Explique el concepto de prevalencia y cómo ésta afecta el aprendizaje de los algoritmos y por lo tanto su rendimiento.

Rta:

En un problema con prevalencia del 3%, solo 3 de cada 100 ejemplos corresponden a la clase positiva. Impacto en el aprendizaje: Los algoritmos de clasificación suelen buscar minimizar un error global, lo que puede llevar a ignorar la clase minoritaria. En contextos de baja prevalencia, los modelos tienden a favorecer la clase mayoritaria para obtener mayor precisión global (accuracy), pero con un alto costo de falsos negativos. Esto puede generar modelos sesgados, que tienen un recall bajo para la clase minoritaria (no detectan bien los casos positivos).

* 1. ¿Por qué la Exactitud (*Accuracy*) deja de ser una métrica confiable en problemas con niveles de prevalencia demasiado bajos, por ejemplo, del 3%?

Rta:

En casos de baja prevalencia, un modelo puede alcanzar una alta accuracy simplemente prediciendo siempre la clase mayoritaria. Por ejemplo, si el modelo siempre predice "negativo", acertará el 97% de las veces (porque solo el 3% son positivos). Esto da un accuracy del 97%, aunque el modelo no detecta ni un solo caso positivo, lo que puede ser inaceptable en muchos contextos (como medicina, seguridad, etc.). Por eso, en problemas desbalanceados se prefieren métricas como: Precision, Recall, F1-score y AUC-ROC

1. Las matrices de confusión o tablas de contingencia resultan ser una herramienta bastante útil a la hora de evaluar clasificadores. Responda las siguientes preguntas, en términos bayesianos o de probabilidades condicionales, basándose en la siguiente matriz:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Actual | | |
| Moroso | Al Día | TOTAL |
| Predicción | Moroso | 400 | 600 | 1000 |
| Al Día | 1000 | 1500 | 2500 |
| TOTAL | 1400 | 2100 | 3500 |

* 1. Reporte la precisión de este clasificador. Interprete

Rta:

La precisión se define como:

De todas las personas que el clasificador predijo como morosas, solo el 40% lo son realmente. Esto indica una alta tasa de falsos positivos (600 de 1000), por lo tanto la confianza en una predicción de “moroso” es baja

* 1. Reporte la exactitud (*accuracy*) de este clasificador. Interprete

Rta:

La exactitud se define como:

El clasificador acierta el 54.3% de las veces, lo cual es apenas mejor que un clasificador aleatorio, sobre todo considerando que hay un desbalance en la proporción de clases (más “al día” que “morosos”)

* 1. Basado en la interpretación de los dos literales anteriores, ¿el clasificador es un buen clasificador? ¿Necesita más información para concluir? ¿Cuál? Explique

Rta:

No es un buen clasificador. La precisión baja (0.40) indica que muchas personas fueron mal clasificadas como morosas. La exactitud moderada (0.543) puede ser engañosa si hay desbalance en clases.

* 1. Interprete el valor resultante de dividir la precisión del clasificador entre la proporción real de morosos [i.e *Precisión/P(Moroso = 1)*]

Rta:

Este cociente mide cuán informativa es la predicción de “moroso” en relación con simplemente suponer que alguien es moroso al azar. Un valor de 1 implica que predecir "moroso" no da ninguna ventaja informativa sobre asumir aleatoriamente que alguien lo es con base en la prevalencia. Es como si el modelo no tuviera poder predictivo

* 1. Basado en su interpretación dada en el literal inmediatamente anterior, ¿el clasificador es un buen clasificador? Explique

Rta:

No, Un cociente de 1 sugiere que la predicción no agrega información útil más allá de la prevalencia base. El modelo no diferencia eficazmente entre morosos y no morosos. Para ser útil, ese cociente debería ser significativamente mayor a 1.

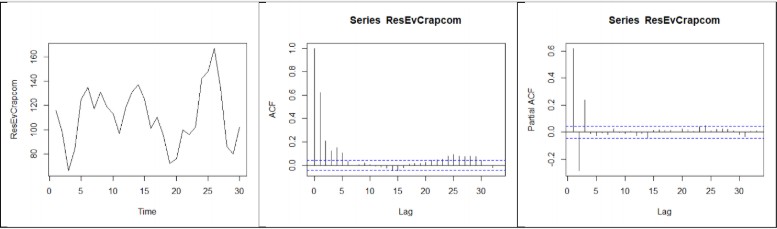
* 1. ¿Qué preguntas le haría a su colega acerca de la construcción de esta matriz de confusión? Enumere máximo tres (3) siendo la primera la más importante y la tercera la menos importante según su criterio.

Rta:

¿El conjunto de datos estaba balanceado o se aplicaron técnicas de oversampling/undersampling?

¿Se evaluó el clasificador en datos no vistos (test set o validación cruzada)?

q¿Qué variable o conjunto de características utilizó el modelo para tomar la decisión de clasificar como "moroso" o "al día"?

1. La empresa CRAPCOM, desarrolladora de videojuegos, desea analizar para uno de sus juegos más populares, cómo ha cambiado en los últimos meses el total de horas jugadas por sus usuarios. Los resultados semanales, en miles de horas, se presentan a continuación.
   1. ¿Considera que la serie es estacionaria?

Rta:

Una serie es estacionaria si sus propiedades estadísticas como la media, la varianza y la autocorrelación se mantienen constantes en el tiempo. Se observa una tendencia leve y posibles ciclos, lo cual sugiere que la serie podría no ser estrictamente estacionaria. Sin embargo, no hay una tendencia creciente o decreciente clara, lo que puede indicar cuasi-estacionariedad.

* 1. Plantee un posible modelo ARMA/ARIMA/SARIMA teniendo en cuenta el ACF y PACF

1. Sea el modelo estacionario 𝑥𝑡 = 𝑐 + 𝑎1𝑥𝑡−1 + 𝑎2𝑥𝑡−2 + ϵ𝑡. El valor esperado 𝐸(∆ 𝑥𝑡) es:
   1. 𝑐/(𝑎1 + 𝑎2)
   2. 𝑐 𝑎1 𝑎2/(1 − 𝑎1 − 𝑎2)
   3. 𝑐/(1 − 𝑎1 − 𝑎2)
   4. 𝑐 + 𝑎1 + 𝑎2
   5. 𝑐
   6. 0
   7. 𝑁𝑖𝑛𝑔𝑢𝑛𝑎
2. Utilizando el conjunto de datos proporcionado en la carpeta comprimida ZIP realice un notebook en Python o R donde haga un resumen que contenga la siguiente información:

Rta: se les da respuesta en el notebook con todo el procedimiento

* 1. Análisis descriptivo de las variables manner\_of\_death y armed utilizando características sociodemográficas (age, gender, race, etc)
  2. Evolución en el tiempo de las muertes por estado (Seleccione los 5 estados con más muertes para facilitar la visualización)
  3. Realice un análisis del ingreso, escolaridad, raza y pobreza para cada uno de los 5 estados

1. Utilizando el mismo conjunto de datos del punto 8, implemente un modelo de aprendizaje automático que dadas ciertas variables sociodemográficas y del tipo de asesinato policiaco, permita predecir la raza de la persona

Rta: se les da respuesta en el notebook con todo el procedimiento